1. Prompt + Target Token Selection

- Prompt text box: for users to paste or type the input prompt.
- Token preview area: displays the tokenization result as individual token chips/spans.
- Target token selection (two supported methods):
  - Click-to-select: user clicks a token chip to set it as the target token.
  - Index input: user enters a numeric token index to select the target token.

Target confirmation UI:

- Clearly show the current selection: Target token and Target index.

2. Token Heatmap

- Tokens are rendered as a sequence of inline spans/chips matching the token preview order.
- Each token's background intensity is driven by its attribution score:
  - Higher score -> stronger highlight
  - Lower score -> lighter/no highlight
- Hover tooltip on each token displays:
  - Raw attribution value
  - Normalized percentage

3. Ranked List (Top-K Most Influential Tokens)

- A side panel listing the Top-K tokens by attribution score.
- Each row includes:
  - Token text
  - Score
- List is sorted highest -> lowest.
- Click interaction:
  - Clicking a row highlights (and optionally scrolls to) the corresponding token in the prompt/heatmap view.

4. Export

- Export JSON: exports the attribution output in a structured format, including:
  - tokens
  - target token/index
  - attribution scores
- Export PNG
  - Exports a snapshot of the heatmap visualization as an image.

5. Request Status & Queue Feedback

- Because the GPU backend may be busy, the frontend must provide clear runtime feedback:
  - Loading state (request submitted, computation in progress)
  - Queue / waiting state (request is queued; show waiting prompt)

- Error / timeout state
  - Display a user-friendly message (timeout or backend error)
  - Provide a simple Retry option using the same prompt/target selection